June 1946

BIOMETRICS

Vol. 2 No. 3 BULLETIN

THE BIOMETRICS SECTION, AMERICAN STATISTICAL ASSOCIATION

MISSING-PLOT TECHNIQUES

R. L. ANDERSON
Institute of Statistics, University of North Carolina

In experimental work it frequently happens that one or more experimental units is missing from the data or has to be rejected because of conditions outside the control of the experimenter. It should be cautioned that observations should be rejected in the analysis of results only under extreme circumstances, when it is quite obvious that the treatment being studied is not responsible for the apparently anomalous results.

One of the first papers on the subject of estimating the yield of a missing unit in field experimental work was published by Allan and Wishart (1). They derived formulas and illustrated their use for a single missing plot in a randomized block and in a Latin square experiment. These methods were extended by Yates (7) to cover several missing units in a given experiment.

The formula given by Yates for estimating the yield of a single missing unit in a randomized block experiment is

$$y=\frac{(rB+tT-G)}{(r-1)(t-1)}$$

where r = the number of blocks and t = the number of treatments in the experiment, B = the total yield of the remaining units in the block where the missing unit appears, T = the total of the yields of the treatment with the missing unit, and G = the grand total.

Similarly for a single missing unit in a Latin square,

$$y=\frac{r(R+C+T)-2G}{(r-1)(r-2)}$$

where r = the number of rows, columns and treatments, and R and C represent the total yields of the remaining units in the row and column in which the missing unit appears.

Yates uses these formulas for several missing units by means of iterative methods, giving an example for an 8x10 randomized block experiment with 9 missing units.

He also shows that in a complete analysis of variance using the above missing-plot values, the treatment sum of squares is over-estimated but may be corrected by subtracting the bias. The bias in a randomized block experiment with one missing unit is

$$\frac{1}{t(t-1)}[B-(t-1)y]^2.$$

With one missing unit of a Latin square, the bias in the treatment sum of squares is

$$\frac{1}{(r-1)^2(r-2)^2}[G-R-C-(r-1)T]^2.$$

Yates gives a generalized formula for the bias for any number of missing units in a randomized block experiment. In practice, this

*A square lattice has k^2 treatments. An orthogonal lattice is one for which the separate groups are orthogonal.

treatment bias is so small that it can be neglected.

Finally Yates presents formulas for the variances of treatment means with missing units. The difference between the mean of a treatment with the yield of a missing unit estimated by the above methods and one with no missing units has a variance in a randomized block experiment of

$$\frac{\sigma^2}{r}\left[2+\frac{t}{(t-1)(r-1)}\right]$$

and in a Latin square experiment of

$$\frac{\sigma^2}{r}\left[2+\frac{r}{(r-1)(r-2)}\right],$$

where σ^2 is the variance of the yield of a single unit. σ^2 is estimated by s^2.

When both treatments in a comparison involve one or more missing units, the formulas for variances between means are more complex because of a correlation between the means. Yates presents an approximate method for handling this case.

In subsequent articles, Yates (8,9) describes methods of handling Latin square experiments in which whole rows, columns or treatments are missing, or in which one row and one column or one row (or column) and one treatment are missing.

The problem of missing-plots becomes much more complicated when we consider the various incomplete block designs. Cornish has described missing-plot formulas for incomplete randomized block designs, simple and triple square lattices, cubic lattices and lattice squares, when the intra-block variance alone was used in estimating the error variance (4). Bliss presents an example of the use of the above missing-plot formula for lattice squares in some corn selection tests (2).

The missing-plot formulas for any orthogonal square lattice* may be generalized as:

$$y=\frac{k^2(r-1)T+rdk(r-1)B-rk(V_2+V_3+\ldots)-k(r-1)U_1+k(U_2+U_3+\ldots)-rS_1+G}{(r-1)(k-1)(rdk-k-1)}$$

where the missing unit occurs in group 1** (interchange numbers if in another group.) We assume d duplications of an r-group lattice ($r=2$ for a simple lattice and $r=3$ for a triple lattice) with k units per block. T and B are the respective total yields for the treatment and block with the missing unit, G is the grand total, and S_1 is the total for group 1. V_2 is the total yield in *group 2 only* of all treatments appearing in the same 2-block as the treatment with the missing unit (including this treatment); U_2 is the total yield for the *entire* experiment of the same treatments as V_2. Similarly for 1, 3, . . . These totals refer to the following numbers of units: $T(rd-1)$, $B(k-1)$, $V_i(dk)$, $U_i(rdk-1)$, $U_i(rdk-1)$, $S_i(dk^2-1)$, $G(rdk^2-1)$ for $i=2, 3, \ldots$.

Cornish (5) later published two articles on methods of handling lattice square experiments with missing-plots when the inter-block information is also used in estimating the error variance. The computations required in these analyses are quite complex. We may summarize the procedure for the square lattices as follows:

(1) Calculate the true intra-block sum of squares using the intra-block estimate of the missing unit y given above. Call this S_E.

(2) Calculate the sum of squares for the group by treatment inter-action using a completely randomized block analysis with the randomized block estimate of y. Call this sum of squares S (rand).

(3) Subtract S (rand) from the total sums of squares corrected for the general mean of the existing values. This gives an unbiased estimate of the reduction in sum of squares due to groups and treatments. Call this S_{GT}.

(4) Subtract S (rand) from the corrected total sum of squares of existing values. This gives the reduction in sum of squares due to groups, treatments and blocks. Call this S_{GTB}.

(5) Subtract S_{GT} from S_{GTB} to give inter-block sum of squares, eliminating treatments.

(6) If only a few of the units are missing, the usual formulas for the weights w and w' can be used to estimate the block adjustments. Cornish states that if as many as 10% of the units are missing, it is doubtful if the experiment is worth analyzing except in very special cases. If an experiment with many missing-

**Group 1 is usually denoted as Group X, Group 2 as Group Y, etc.

plots is analyzed, a method of adjusting the formulas for w and w' must be derived. Cornish indicates the general attack on this problem.

It appears that the randomized block formula can be used to estimate the yields of missing units for the lattice designs without sacrificing much information. A study should be made of the average relative bias in the error variance when the randomized block estimate of the yield of a missing unit is used in lattice designs.

The problem of missing units in confounded factorial experiments has been studied by Cochran. When one or several factors in the 2^n and 3^n designs are completely confounded (hence, the same design is duplicated in all replications), the formula for the yield of a missing unit is

$$y = \frac{rB + kT - B'}{(r-1)(k-1)}$$

where r= the number of replications, k= the number of units per block, B= the total yield of the other *(k—1)* units in the same block as the missing unit, B'= the total yield of all *r* blocks having the same set of treatments as the block with the missing unit (including this block), and T= the total yield of the *(r—1)* other units which have the same treatment as the missing unit. Cochran has also developed a formula which holds for partial confounding when (a) a different replication of the basic plan is used for each replication of the experiment and (b) no treatment comparison is confounded in more than one replication. These results are included in a mimeographed set of notes by Cochran and Cox on Design of Experiments (3).

If high-order interactions are used as estimates of error, missing values should be determined by the process of minimizing the error variance. The basic method used in obtaining all of these missing-plot formulas is to let y represent the missing value, set up the sum of squares for error in terms of y, and then estimate y by minimizing this error variance. This method is also explained by R. A. Fisher (6) with an example of its application to a 6x6 Latin square. In certain complicated cases, it may be better to use a complete least-square solution or to use the covariance technique which will be outlined below for missing units in split-plot experiments. Yates outlines the method of fitting constants by least squares in his article on Latin squares (8,9).

Split-plot experiments. The present author derived some formulas for missing plots in split-plot experiments by minimizing the error variance. The method of covariance will be used in the derivations which follow in order to furnish an easy means for estimating the bias in the treatment sum of squares. We shall assume that we have a split-plot experiment with r replication and α whole-plot and β sub-plot treatments so that the total number of units is $N = r\alpha\beta$.

One sub-plot missing. Let the missing unit be that for the whole-plot treatment a_1, sub-plot treatment b_1 and replication r_1. Also let A_1 be the total yield of all existing units with treatment a_1, B_1 the total yield of all existing units with treatment b_1, R_1 the total yield in replication r_1, (A_1B_1) the total yield of all existing units with both a_1 and b_1, (R_1A_1) the total yield of all existing units with both r_1 and a_1 and G the grand total. Let x=0 and y= the actual yield for the existing units and x=—1 and y=0 for the missing unit.

In the analysis of covariance table, $S(x^2)$ equals the degrees of freedom divided by N in all cases. The cross-product sums $S(xy)$ are given in Table 1.

The best estimate of the yield of the missing unit in order to minimize the sub-plot error is simply the error b regression coefficient,

$$y = \frac{r(R_1A_1) + \beta(A_1B_1) - A_1}{(r-1)(\beta-1)}$$

If this yield is used for the missing unit, all sums of squares except that for error b will be slightly over-estimated. The unbiased estimate of any sum of squares is found by first computing a new line in Table 1, which is the degrees of freedom and $S(xy)$ for this sum plus error b. The new $S(x^2)$, which is the degrees of freedom divided by N, and the new $S(xy)$ are designated as $S(x^2_1)$ and $S(x_1y)$. Then the new regression coefficient is

$$y_1 = \frac{S(x_1y)}{S(x_1)^2}$$

Table 1

Sums of Cross-products for Split-plot Experiment with one Missing Unit.

Replications	r−1	$-\frac{R_1}{a\beta}+\frac{G}{N}$
Treatment A	a−1	$-\frac{A_1}{r\beta}+\frac{G}{N}$
Error *a* (E_a)	(r−1) (a−1)	$\frac{-ra(R_1A_1)+rR_1+aA_1-G}{N}$
Treatment B	β−1	$-\frac{B_1}{ra}+\frac{G}{N}$
A×B	(a−1) (β−1)	$-\frac{a\beta(A_1B_1)+aA_1+\beta B_1-G}{N}$
Error *b* (E_b)	a(r−1) (β−1)	$\frac{ra(R_1A_1)+a\beta(A_1B_1)-aA_1}{N}$

The bias in estimating the sum of squares under consideration is

$$(y-y_1)^2S(x_1)^2$$

Note that the bias is always positive; that is, the sum of squares is always over-estimated in the analysis of variance.

Thus, for the treatment B,

$$y_1=\frac{ra(R_1A_1)+a\beta(A_1B_1)-aA_1-\beta B_1+G}{(\beta-1)\ (ra-a+1)}$$

and $S(x_1^2)=(\beta-1)\ (ra-a-1)\ /ra\beta$.

For the interaction AB,

$$y_1=\frac{ra(R_1A_1)+\beta B_1-G}{(ra-1)\ (\beta-1)}$$

and $S(x_1^2)=(ra-1)\ (\beta-1)\ /ra\beta$.

An exact treatment of the whole-plot analysis is somewhat more complicated. It is unlikely that the test of significance for treatment A would be seriously affected by the slight biases introduced by use of the missing-plot value. One possibility of obtaining more exact estimates of the sums of squares for treatment A and for error *a* would be to minimize the sum of squares for error *a* and calculate the true sum of squares for treatment A on this basis. If this were done, the estimate of the missing value would be

$$y'=-\frac{ra(R_1A_1)-rR_1-aA_1+G}{(r-1)\ (a-1)}$$

However, this gives the same result as would be obtained by considering the entire whole plot (R_1A_1) to be missing, which does not seem justified. A second method would be to use the above covariance technique for B and AB to obtain unbiased estimates of the sum of squares for both A and error *a*. However, these two adjusted sums of squares would no longer be independent and the F-test could not be used to test the significance of the A differences.

As an example of the missing-plot technique with a single missing unit, consider the following wheat straw yields from 4 top dressings of nitrogen (0, 15, 30 and 45 pounds) each replicated 4 times, constituting the 16 whole plots, and 4 types of fertilizers on the subplots (0-0-0, 15-0-0, 0-40-40, and 15-40-40.) The experiment was conducted by the N. C. Agricultural Experiment Station as part of a Cooperative Fertilizer Experiment. The yields are grams of straw per plot.

Using the missing plot formula,

$$y=\frac{4(1930)+4(2056)-9080}{9}=763$$

The analysis of variance for these data, using y=763 and without correction for bias, is presented in Table 3.

Since the treatment B effect is highly significant (and far beyond the 1% point) and the AB interaction is definitely non-significant, we need not be concerned with the bias in the estimates of the sums of squares. For purposes of illustration, we have computed the actual bias in each sum of squares. For B, the bias is 18,238 (a 7% bias); and for A×B the bias is 340 (1%). These would be subtracted from the values in Table 3 for unbiased estimates.

Table 2.

Yields of Wheat Straw in a Split-plot Experiment.

lb. N Top Dress	Fertilizer Type of	Replications I	II	III	IV	Total	
0	0- 0- 0	332	260	202	210	1004	
	15- 0- 0	346	334	232	228	1140	
	0-40-40	340	236	250	217	1043	
	15-40-40	454	476	346	384	1660	
	Total	1472	1306	1030	1039	4847	
15	0- 0- 0	412	384	362	348	1506	
	15- 0- 0	420	606	366	426	1818	
	0-40-40	540	410	380	604	1934	
	15-40-40	604	522	508	510	2144	
	Total	1976	1922	1616	1888	7402	
30	0- 0- 0	542	472	516	458	1988	
	15- 0- 0	638	572	652	550	2412	
	0-40-40	693	538	434	614	2279	
	15-40-40	844	708	744	706	3002	
	Total	2717	2290	2346	2328	9681	
45	0- 0- 0	730	590	294	560	2174	
	15- 0- 0	664	616	418	702	2400	
	0-40-40	740	724	454	532	2450	
	15-40-40	738	y	650	668	2056	$=(A_1B_1)$
	Total	2872	1930= (R_1A_1)	1816	2462	9080	$=A_1$
Total		9037	7448=R_1	6808	7717	31010	=G

B_1=1660+2144+3002+2056=8862

Table 3.

Analysis of Variance of Wheat Straw Data.

	Degrees of Freedom	Sum of Squares	Mean Square
Whole plots			
Replications	3	162,998	54,333
Top Dressing (A)	3	1,031,784	343,928**
Error a	9	80,644	8,960=E_a
Sub-plots			
Seedings (B)	3	283,167	94,389**
A×B	9	29,391	3,266
Error b	35	151,950	4,341=E_b

Entered as second-class matter, May 25, 1945, at the post office at Washington, D. C., under the Act of March 3, 1879. The Biometrics Bulletin is published six times a year—in February, April, June, August, October and December—by the American Statistical Association for its Biometrics Section. Editorial Office: 1603 K Street, N.W., Washington 6, D. C.

Membership dues in the American Statistical Association are $5.00 a year, of which $3.00 is for a year's subscription to the Quarterly Journal, fifty cents is for a year's subscription to the ASA Bulletin and members who pay $1.00 additional receive a year's subscription to the Biometrics Bulletin. Dues for Associate members of the Biometrics Section are $2.00 a year, of which $1.00 is for a year's subscription to the Biometrics Bulletin. Single copies of the Biometrics Bulletin are 60 cents each and annual subscriptions are $2.00 Subscriptions and applications for membership should be sent to the American Statistical Association, 1603 K Street, N.W., Washington 6, D. C.

Similarly the A treatments are definitely different regardless of any slight bias. The bias, as computed by the second method, in the sum of squares for error a is 2531 (3%), and for treatment A is 23,172 (2%).

If it is desired to test the significance of the difference between two mean yields, one with and one without a missing unit, the usual formulas for the variance between two means do not hold. The following formulas give the variances of the difference between two such means (assume a_2 and b_2 represent treatments with no missing units.)

$$\bar{b}_1-\bar{b}_2: \frac{2E_b}{r\alpha}\left[1+\frac{\beta}{2\alpha(r-1(\beta-1)}\right] = \frac{8682}{16}\left[1+\frac{4}{72}\right]=573$$

$$\bar{a}_1-a_2: \frac{2}{r\beta}\left[E_a+\frac{E_b}{2(r-1)(\beta-1)}\right] = \frac{2}{16}\left[8960+\frac{4341}{18}\right]=1150$$

$$\overline{a_1b_1}-\overline{a_1b_2}: \frac{2E_b}{r}\left[1+\frac{\beta}{2(r-1)(\beta-1)}\right] = \frac{8682}{4}\left[1+\frac{4}{18}\right]=2653$$

$$\overline{a_1b_1}-\overline{a_2b_1}: \frac{2E_a}{r\beta}+\frac{2E_b}{r\beta}\left[(\beta-1)+\frac{\beta^2}{2(r-1)(\beta-1)}\right] = \frac{17920}{16}+\frac{8682}{16}\left(3+\frac{16}{18}\right)=3230$$

In these formulas, the unbiased E_a should have been used; however, the bias is usually so small that the difference would not be important. For example, in the above, E_a would have been 8679 instead of 8960. Hence the variance of $(\bar{a}_1-a_2)=1115$. The standard error of this difference would be 33.4 instead of 33.9.

One whole-plot missing. The whole-plot analysis can be made on a randomized block basis (see page 41), using the same missing-plot formula. The treatment A bias and the variance of the difference between two treatment means must be divided by β, if the results are put on a sub-plot basis.

For the sub-plots, the method of proportionate sub-class numbers can be used to evaluate the B and AB sums of squares. The sub-plot error (E_b) can be obtained by subtracting the variation between existing whole-plots plus the B and AB variations from the total variation between existing plots. In other words, the analysis of the sub-plots need not concern itself with the missing units. In terms of the notation given in the previous section for one missing sub-plot, the various sums of squares are as follows:

$$B: \frac{SB^2}{r\alpha-1}-\frac{G^2}{\beta(r\alpha-1)}$$

$$AB: \left\{\frac{S(A_iB)^2}{r-1}+\frac{\sum_{i=2}^{\alpha} S(A_iB)^2}{r}\right\}-\left\{\frac{A_1^2}{\beta(r-1)}+\frac{\sum_{i=2}^{\alpha} A_i^2}{r\beta}\right\}-\frac{SB^2}{r\alpha-1}+\frac{G^2}{\beta(r\alpha-1)}$$

$$E_b: Sy^2-\frac{S(RA)^2}{\beta}-(\text{sum of squares for B and AB.})$$

The variances of the differences between sub-plot treatment means are as follows:

$$\bar{b}_1-\bar{b}_2: \frac{2E_b}{r\alpha-1}$$

$$\overline{a_ib_i}-\overline{a_ib_2}: \frac{2E_b}{r-1}$$

$$\overline{a_2b_1}-\overline{a_2b_2}: \frac{2E_b}{r}$$

$$\overline{a_ib_1}-\overline{a_2b_1}: \left(\frac{E_a+(\beta-1)E_b}{\beta}\right)\left(\frac{1}{r}+\frac{1}{r-1}\right)$$

$$\overline{a_2b_1}-\overline{a_2b_1}: \frac{2}{r}\left(\frac{E_a+(\beta-1)E_b}{\beta}\right)$$

REFERENCES CITED

1. Allan, F. E. and J. Wishart. A method of estimating the yield of a missing plot in field experimental work. Jour. Agr. Sci. 20, Pt. 3, 399-406. 1930a.
2. Bliss, C. I. and R. B. Dearborn. The efficiency of lattice squares in corn selection tests in New England and Pennsylvania. Proc. Am. Soc. Hort. Sci. 41:324. 1942.
3. Cochran, W. G. and Gertrude Cox. Experimental designs. (Mimeographed.) Institute of Statistics, Raleigh, N. C.
4. Cornish, E. A. The estimation of missing values in incomplete randomized blocks experiments. Ann. Eugen. 10: 112-118. 1940.
   Cornish, E. A. The estimation of missing values in quasi-factorial designs. Ibid. 10:137-143. 1940b.
   Cornish, E. A. The analysis of quasi-factorial designs with incomplete data. 1. Incomplete Randomized Bl. Jour. Aust. Inst. Agr. Sci. 6:31-39. 1940c.
   Cornish, E. A. The analysis of quasi-factorial designs with incomplete data. 2. Lattice Squares. Ibid. 7:19-26. 1941a.
   Cornish, E. A. The analysis of quasi-factorial designs with incomplete data. 3. Square, Triple, and Cubic Lattices. (Unpublished) 1941b.
5. Cornish, E. A. The recovery of inter-block information in quasi-factorial designs with incomplete data. 1. Square, Triple, and Cubic Lattices. Bul. 158 (1943). 2. Lattice Squares, Bul. 175. 1944. Coun. Sci. Ind. Res. (Aust.)
6. Fisher. R. A. The design of experiments. Second edition. Oliver and Boyd, Edinburgh. Sec. 58. 1. 1937.
7. Yates, F. The analysis of replicated experiments when the fields results are incomplete. Emp. Jour. Exp. Agr. 1:129-142. 1933.
8. Yates, F. Incomplete Latin Squares. Jour. Agri. Sci. 26 Pt. 2. 301-315. 1936.
9. Yates, F. and R. W. Hale. The analysis of Latin squares when two or more rows. columns or treatments are missing. Supp. Jour. Royal Stat. Soc. 6: No. 1:67-79. 1939.

LIMITATIONS OF THE APPLICATION OF FOURFOLD TABLE ANALYSIS TO HOSPITAL DATA*

JOSEPH BERKSON, M.D.,
Division of Biometry and Medical Statistics, Mayo Clinic,
Rochester, Minnesota

In the biologic laboratory we have a method of procedure for determining the effect of an agent or process that may be considered typical. It consists in dividing a group of animals into two cohorts, one considered the "experimental group," the other the "control." On the experimental group some variable is brought to play; the control is left alone. The results are set up as in table 1-a. If the results show that the ratio $a:a+b$ is different from the ratio $c:c+d$, it is considered demonstrated that the process brought to bear on the experimental group has had a significant effect.

A similar method is prevalent in statistical practice, which I venture to think has come into authority because of its apparent equivalence to the experimental procedure. In Biometrika it is referred to as the fourfold table and it is used as a paradigm of statistical analysis. The usual arrangement is that given in table 1-b. The entries, a, b, c and d are manipulated arithmetically to determine whether there is any correlation between A and B. A considerable number of indices have been elaborated to measure this correlation. Pearson has given the formula for calculating the product-moment correlation coefficient from a fourfold table on the assumption that the distribution of both variates is normal; Yule has an index of association for the fourfold table; there are the chi-square test and others. In essence, however, all these indices measure in different ways whether and how much, in comparison with the variation of random sampling, the ratio $a:a+b$ differs from the ratio $c:c+d$. If the difference departs significantly from zero, there is said to be correlation, and the correlation is *the* greater the greater the difference.

Now there is a distinction between the method as used in the laboratory and as

*This paper was presented in somewhat different form at a meeting of the American Statistical Association in 1938. Recent inquiries have prompted its publication at this time.

applied in practical statistics. In the experimental situation, the groups, B and not B, are selected *before* the subgroupings, A and not A, are effected; that is, we start with a total group of unaffected animals. In the statistical application, the groupings, B and not B, are made *after* the subgroupings, A and not A, are already determined; that is, all the effects are already produced *before* the investigation starts. In the end, the tables of the results which are drawn up *look* alike for the two cases, but they have been arrived at differently. Correlative to this difference, a different interpretation may apply to the results, and this paper deals with a specific case of a kind that arises frequently in a medical clinic or a hospital. I take an example.

There was prevalent an impression that cholecystic disease is a provocative agent in the causation or aggravation of diabetes. In certain medical circles, the gall bladder was being removed as a treatment for diabetes. The authorities of a hospital wish to know whether their accumulated records of incidence, examined statistically, support this practice. On the face of it, it would appear that we have here the typical and elementary problem of the comparision of rates in a fourfold table. The total population of patients for a period is to be divided into two groups, "diabetes" and "no diabetes" and the rate of incidence of cholecystitis in the one compared with the rate in the other. Accordingly, table 2 was set up.

Table 2 shows a significant difference indicating positive correlation between cholecystitis and diabetes. An objection which might be brought against this particular tabulation is that the "not diabetes" group consisting as it does of all patients without diabetes, will contain a variety of diagnoses, some of which may themselves be correlated with cholecystitis, even as diabetes may be; hence the control may be considered not good. To meet this objection we do not select for the control group the entire nondiabetic population, but take a diagnosis which cannot reasonably be thought to be correlated with cholecystitis and use this as a criterion for the control group. I took, in fact, several refractive errors of the sort for which patients come to the clinic for glasses as such a diagnostic group, and table 3 was the result.

Again we see that the difference is positive and significant in comparison with the probable error, and the usual judgment would be that cholecystitis and diabetes are positively correlated. Of course, in any detailed analysis we should wish to keep age and sex constant, inquire into the reliability of the diagnoses, and so forth. But the point referred to in this paper has no relation to such questions, and for the sake of the argument we shall consider that all such factors have been adequately controlled. Even so, do the results permit any conclusion as to whether cholecystitis is biologically correlated with diabetes?

Since the hospital population comes from the general population, let us begin there. For the sake of simplification, we shall consider only the three diseases referred to, cholecystitis, diabetes and refractive errors. If the incidence of these conditions in the general population is represented by p_c, p_d and p_r and there is no correlation between the diseases, we have for the constitution of the population the expressions shown in table 4, in which n_d is the number having diabetes but not having cholecystitis nor refractive errors, n_{dc} those having diabetes and cholecystitis but not having refractive errors, n_{dcr} those having diabetes, cholecystitis and refractive errors, n_0 those having none of these diseases, and so forth. N is the total population. If we assume for illustrative purposes, a population of 10,000,000 persons, and $p_d = 0.01$, $p_c = 0.03$, and $p_r = 0.10$, the numbers of the various constituents are given in table 4. From these figures we may set up two fourfold tables as before (table 5).

In both parts of table 5 it is seen that the difference of the pertinent ratios is zero, which is as it should be, since there is no correlation. This result, of course, could have been foreseen without this computation but I desired to establish the numbers for use later. Now suppose we follow that portion of the population which gets to the hospital. For this purpose we must develop some elementary relationships.

We shall suppose that associated with each

particular disease is a definite probability that its victims will be selected for the hospital. That is, we shall suppose that a person who has cholecystitis has a certain definite probability of being drawn to the hospital because of the presence of that disease alone, and so for other diseases. Furthermore, for simplicity we shall say that these selective probabilities operate independently, as though a person who had two diseases were like Siamese twins, each one of whom had one disease, so that the probability of the twins' coming to the hospital is the probability of either one getting there, but the presence of one disease does not affect the other in any way. Let the selective rates be represented by s_1, s_2, s_3, and so forth and their complements $(1 - s)$ be represented by t_1, t_2, t_3, and so forth, the number in the general population by n and the number in the hospital by n'. Then, we have the following equations:

$$n'_1 = n_1(1-t_1) = n_1(s_1)$$
$$n'_{12} = n_{12}(1-t_1t_2) = n_{12}(s_1+s_2-s_1s_2)$$
$$n'_{123} = n_{123}(1-t_1t_2t_3) = n_{123}(s_1+s_2+s_3-s_1s_2-s_1s_3-s_2s_3+s_1s_2s_3)$$

From these relationships an interesting conclusion can at once be drawn. Suppose all the s's are equal, but small; then the following ratios will result:

$$\frac{n'_{12}}{n'_1} = \frac{n_{12}}{n_1}(2-s) = \text{approximately, } \frac{n_{12}}{n_1} \times 2$$

$$\frac{n'_{123}}{n'_1} = \frac{n_{123}}{n_1}(3-3s+s^2) = \text{approximately, } \frac{n_{123}}{n_1} \times 3$$

From these equations it is seen that the ratio of multiple diagnoses to single diagnoses in the hospital will always be greater than in the general population; for two diagnoses the ratio will be about twice that of the general population, for three diagnoses about three times, and so forth.

Let us now apply the appropriate factors of selection to the various constituents of the hypothetical general population which have been enumerated. Assuming as a simple instance that all the selective probabilities are equal and have the value 0.05, the frequencies given in tables 6 and 7 will result.

We see here that though in the general population, the incidence of cholecystitis was identical among the persons who had diabetes and the persons who had refractive errors, in the hospital population the incidence was less in the diabetic group than in the control group, giving an appearance of a small negative correlation, and this in the face of the fact that we have assumed equality of selective rates for the various diseases.

In general the selective rates can be assumed to be anything but equal for different diseases. Various circumstances, such as the severity of the symptoms, the amenability of the disease to treatment by a local physician or the reputation of a particular hospital for treatment of particular diseases, will determine the probability that a specific disease will bring its victim to a particular hospital. To see the effect of a variation in selective rates, let us hypothesize some values which will differ among themselves as follows: $s_c = 0.15$, $s_d = 0.05$, $s_r = 0.20$. The resulting numbers of the various constituents of the population that will come into the hospital are shown in table 8 and the fourfold table drawn up from these figures is given as table 9.

We now find that the incidence of cholecystitis in the diabetic group is about twice that of the control. This would show, so far as the hospital population is concerned, a positive correlation between cholecystitis and diabetes, but it would be quite unrepresentative of the situation in the general population and of no biologic significance.

The relationships dealt with arithmetically in the previous tables are given algebraically as follows:

$$p'_{1.2} = \frac{p_1q_3(1-t_1t_2)+p_1p_2(1-t_1t_2t_3)}{p_1q_3(1-t_1t_2)+q_1q_3(1-t_2)+p_1p_3(1-t_1t_2t_3)+p_2q_1(1-t_2t_3)}$$

$$p'_{1.3} = \frac{p_1(1-t_1t_3)}{p_1(1-t_1t_3)+q_1(1-t_3)}$$

Where

$p'_{1.2}$ is the incidence in the hospital population of condition 1 among persons who have condition 2

$p'_{1.3}$ is the incidence in the hospital population of condition 1 in the control group who have condition 3

p_1, p_2, and p_3 are the independent probabilities in the general population of conditions 1, 2 and 3, $q = 1 - p$

t_1, t_2, and t_3 are the complements $(1-s)$ of the independent selective probabilities s_1, s_2 and s_3 applying to condition 1, 2 and 3

Comment

The assumption made in the text that a probability can be assigned to every disease, which gives the chance that a patient suffering from that disease alone, will come to the hospital is, I think, in general accord with the actual mechanism by which such a patient is selected for the hospital population. The assumption that these probabilities operated independently in an individual who is suffering from more than one disease is doubtless oversimple. In general we may guess that if a patient is suffering from two diseases, each disease is itself aggravated in its symptoms and more likely to be noted by the patient. So far as this difference of fact from assumption goes, its effect would be to increase relatively the representation of multiple diagnoses in the hospital, and in general to increase the discrepancy between hospital and parent population, even more than if the probabilities were independent.

It appears from the development that it is hazardous to apply in a hospital population the method of the fourfold table analysis for an inquiry into the correlation of diseases. This applies also to other similar problems, as for instance whether the incidence of say, heart disease, is different for laborers and farmers, if it is known that laborers and farmers are not represented in the hospital in the proportion that they occur in the community. However, the formulas given indicate some special cases in which comparison is not basically invalid. If the selective rate for any particular condition is zero, the relative incidence of that condition in several disease groups may be validly examined, regardless of the selective rates affecting the other groups. This refers to inquiries in which for instance eye color or anthropologic type is examined in various disease groups to ascertain whether there is correlation between these characters and disease. If each of the disease groups examined consists of only one disease, for example, diabetes or refractive errors but not both, and if the selective rates for these two groups do not differ appreciably then also it is valid to compare the incidence in them of cholecystitis, even though the latter disease is not fairly represented in the hospital.

Except for such cases there does not appear to be any ready way of correcting the spurious correlation existing in the hospital population by any device that does not involve the acquisition of data which would themselves answer the primary question. For instance the device sometimes used of setting up in the hospital sample a one-to-one control so that both groups examined have the same number of cases and are identical as regards say, age and sex does not touch the difficulties referred

Table 1
Fourfold Tables

a Typical of experimental situation				*b* Statistical form			
Group	Effect	No effect	Total	Group	A	Not A	Total
Experimental	a	b	a+b	B	a	b	a+b
Control	c	d	c+d	Not B	c	d	c+d
Total	a+c	b+d	a+b+c+d	Total	a+c	b+d	a+b+c+d

to here. It is to be emphasized that the spurious correlations referred to are not a consequence of any assumptions regarding biologic forces, or the direct selection of correlated probabilities, but are the result merely of the ordinary compounding of independent probabilities. The same results as shown here would appear if the sampling were applied to randomly distributed cards instead of patients.

Table 2

Relation of cholecystitis to diabetes—hospital population

	A Cholecystitis	Not A Not cholecystitis	Total
B: Diabetes	28	548	576
Not B: Not diabetes	1,326	39,036	40,362
Total	1,354	39,584	40,938
Cholecystitis in diabetic group			4.86%
Cholecystitis in control group (not diabetic)			3.28%
Difference			+1.58%±0.5%

Table 3

Relation of cholecystitis to diabetes—hospital population, refractive errors used as control

	A Cholecystitis	Not A Not cholecystitis	Total
Diabetes	28	548	576
Refractive errors	68	2,606	2,674
Total	96	3,154	3,250
Cholecystitis in diabetic group			4.86%
Cholecystitis in control group (refractive errors)			2.54%
Difference			+2.32%±0.5%

Table 4

Constitution of general population, various diseases

$n_d = p_d q_c q_r \times N = 87{,}300$

$n_c = p_c q_d q_r \times N = 267{,}300$

$n_r = p_r q_d q_c \times N = 960{,}300$

$n_{dc} = p_d p_c q_r \times N = 2{,}700$

$n_{dr} = p_d p_r q_c \times N = 9{,}700$

$n_{cr} = p_c p_r q_d \times N = 29{,}700$

$n_{dcr} = p_d p_c p_r \times N = 300$

$n_0 = q_d q_c q_r \times N = 8{,}642{,}700$

$N = 10{,}000{,}000$

$p_d = 0.01, \quad p_c = 0.03, \quad p_r = 0.10$

$q_d = 0.99, \quad q_c = 0.97, \quad q_r = 0.90$

Table 5

Cholecystitis and diabetes, general population

	Cholecystitis	Not cholecystitis	Total
Diabetes	3,000	97,000	100,000
Not diabetes	297,000	9,603,000	9,900,000
Total	300,000	9,700,000	10,000,000
Cholecystitis in diabetic group			3%
Cholecystitis in control group (nondiabetic)			3%
		Difference	0%

	Cholecystitis	Not cholecystitis	Total
Diabetes	3,000	97,000	100,000
Refractive errors	29,700	960,300	990,000
Total	32,700	1,057,300	1,090,000
Cholecystitis in diabetic group			3%
Cholecystitis in control group (refractive errors)			3%
		Difference	0%

Table 6

Enumeration of hospital population for $s_d=s_c=s_r=0.05$

General population numbers	f*	Hospital population, expected numbers
$n_d = 87{,}300$	0.05	$n'_d = 4{,}365$
$n_c = 267{,}300$	0.05	$n'_c = 13{,}365$
$n_r = 960{,}300$	0.05	$n'_r = 48{,}015$
$n_{dc} = 2{,}700$	0.0975	$n'_{dc} = 263$
$n_{dr} = 9{,}700$	0.0975	$n'_{dr} = 946$
$n_{cr} = 29{,}700$	0.0975	$n'_{cr} = 2{,}896$
$n_{dcr} = 300$	0.142625	$n'_{dcr} = 43$
$n_0 = 8{,}642{,}700$	0	$n'_0 = 0$

*The fraction of the specified individuals which is selected for the hospital under the operation of the selective forces s. It is equal to 1 minus the products of the appropriate t's; for example $f_{dcr}=1-t_dt_ct_r$.

Table 7

Cholecystitis and diabetes, hospital population: expected numbers for $s_c=s_d=s_r=0.05$

	Cholecystitis	Not cholecystitis	Total
Diabetes	306	5,311	5,617
Refractive errors	2,896	48,015	50,911
Total	3,202	53,326	56,528
Cholecystitis in diabetic group			5.45%
Cholecystitis in control group (refractive errors)			5.69%
		Difference	−0.24%

Table 8
Enumeration of a hospital population for $s_c=0.15$, $s_d=0.05$, $s_r=0.20$

General population numbers	f	Hospital population, expected numbers
n_d = 87,300	0.05	n'_d = 4,365
n_c = 267,300	0.15	n'_c = 40,095
n_r = 960,300	0.20	n'_r =192,060
n_{dc} = 2,700	0.1925	n'_{dc} = 520
n_{dr} = 9,700	0.24	n'_{dr} = 2,328
n_{cr} = 29,700	0.32	n'_{cr} = 9,504
n_{dcr}= 300	0.354	n'_{dcr}= 106
n_0 =8,642,700	0	n'_0 = 0

Table 9

Cholecystitis and diabetes, hospital population expected numbers for $s_c=0.15$, $s_d=0.05$, $s_r=0.20$

	Cholecystitis	Not cholecystitis	Total
Diabetes	626	6,693	7,319
Refractive errors	9,504	192,060	201,564
Total	10,130	198,753	208,883
Cholecystitis in diabetic group			8.55%
Cholecystitis in control group (refractive errors)			4.72%
Difference			+3.83%

STANDARD ERRORS OF YIELDS ADJUSTED FOR REGRESSION ON AN INDEPENDENT MEASUREMENT

D. J. Finney
Lecturer in the Design and Analysis of Scientific Experiment, University of Oxford.

The precision of comparisons between mean results for the several treatments of a well-planned experiment can often be increased by application of the analysis of covariance. If y represents the measurement studied on each experimental unit, and x is a second measurement on each unit, itself unaffected by the treatments under test but showing a significant error correlation with y, the regression of y on x is used to adjust the means for each treatment; if the regression of y on x is highly significant, the standard

errors of differences between the adjusted treatment means may be substantially less than for the unadjusted means. For example, in animal feeding trials weight gains may be adjusted so as to take account of differences in initial weight that, providing the assignment of animals to treatments was random, could not be associated with treatments, or crop yields in field-plot tests may with advantage be adjusted for differences in plant density when variations in the latter are considerable but are independent of treatment.

If t treatments are each tested in r-fold replication, in a simple randomized block or Latin square design, and give mean yields $y_u (u=1, 2, \ldots t)$, the means adjusted for regression are

$$y_u' = y_u - b(x_u - \bar{x}),$$

where the x_u are the treatment means for the independent variate and b is the regression coefficient estimated from the error line of the analysis of variance and covariance. Now the variance of a difference between two unadjusted means is

$$V(y_1 - y_2) = \frac{2s^2}{r}, \qquad (1)$$

where s^2 is the error mean square for the analysis of variance of y; this quantity is, of course, the same for every pair of treatments. The variance of the difference between the corresponding adjusted means is

$$V(y_1' - y_2') = s'^2\left(\frac{2}{r} + \frac{(x_1 - x_2)^2}{A}\right) \qquad (2)$$

where s'^2 is the residual error mean square for y after removal of the regression component and A is the error sum of squares in the x analysis. Since this variance depends upon the pair of treatments compared, the presentation of standard errors in summary tables of means is made difficult. Often the second term in the expression on the right of equation (2) is negligible by comparison with the first, so that little fault is committed by assigning a standard error $s'/\sqrt{r}$ to each adjusted mean. This procedure, however, leads to consistent underestimation of standard errors, since the neglected term is necessarily positive, and should therefore not be employed unless all differences of the type $(x_1 - x_2)$ are very small.

A simple modification, which the writer has found useful when separate computation of equation (2) for every interesting comparison seemed unnecessary, is to remove the bias by inserting an average value for $(x_1 - x_2)^2$. The average value of $(x_u - x_v)^2$ over all possible pairs of different treatments can easily be shown to be $2a/r(t-1)$, where a is the sum of squares for "treatments" in the analysis of variance for x. Hence, with this approximation, each treatment mean may be assigned a variance

$$V(y_u') = \frac{s'^2}{r}\left(1 + \frac{a}{A(t-1)}\right), \qquad (3)$$

and twice this amount may be taken as the variance of the difference between any pair of adjusted treatment means. It may further be shown that the variance of the difference between a mean yield for one set of k_1 treatments and another set of k_2, averaged over all possible selections of the k_1 and k_2, is

$$\frac{k_1 + k_2}{k_1 k_2} V(y_u') = V(y_u')\left(\frac{1}{k_1} + \frac{1}{k_2}\right),$$

and is the same as would be obtained by direct and uncritical use of equation (3).

The average efficiency of comparisons between adjusted treatment means relative to comparisons between unadjusted means is

$$E = \frac{s^2}{s'^2\left(1 + \frac{a}{A(t-1)}\right)}$$

which quantity therefore represents the overall gain in information on treatment contrasts through taking account of the independent variate. Providing that treatments have had no effect on x, the expected value of the factor

$$\left(1 + \frac{a}{A(t-1)}\right) \text{ is } \left(1 + \frac{1}{e}\right),$$

where e is the number of error degrees of freedom; this should not normally be used instead of the experimental value, as considerable variations may occur unless both sums of squares, a and A, are based on large numbers of degrees of freedom.

In an experiment of factorial design, the factor $\left(1 + \frac{a}{A(t-1)}\right)$ could be calculated

separately for each main effect and interaction, taking a as the treatment sum of squares and $(t-1)$ as the corresponding number of degrees of freedom, but often an over-all value computed for the total of all treatment sums of squares will be sufficiently good, since, if the treatments are all without effect on x, the mean squares $a/(t-1)$ may be expected to be about the same size; of course, in calculating this sum of squares, allowance must be made for any partial or total confounding of treatment contrasts. For more complicated designs, such as those of the lattice and other incomplete block types, the treatment contrasts are obtained as weighted means of intra- and inter-block comparisons. As Cochran (1940) has said, the two kinds of comparison should be adjusted separately for their own error regressions. If a composite summary table is required, this may be formed as a weighted mean of the two kinds of contrast, each being weighted inversely as the average variance of comparisons adjusted for the regression. Each variance must contain the appropriate factor corresponding to $\left(1+\frac{a}{A(t-1)}\right)$; here a and A must be taken from the parts of the analysis of variance of x for intra- and inter-block comparisons separately, together with $(t-1)$, the treatment degrees of freedom in each part of the analysis. The standard errors of the weighted means will only be approximately correct, on account of the regression adjustments having introduced correlations between intra- and inter-block components, but the degree of approximation will generally be sufficiently good.

REFERENCE

COCHRAN, W. G. (1940). The Analysis of Lattice and Triple Lattice Experiments in Corn Varietal Tests. II. Mathematical Theory. Iowa Agr. Exp. Sta., Res. Bull. 281.

QUERIES

QUERY: May distribution variances be used directly to form a ratio equivalent to an F ratio, such that probability may then be read directly from conventional F tables, in testing the significance of differences between two variances? If so, what are the assumptions involved in, and the limitations imposed upon the use of such a method?

It may be that the answer to the above question will require more specific information as to the particular circumstances of the application. Hearing thresholds are taken in a group of ears before and after exposure to aircraft noise. In addition to differences between means before and after we are interested in knowing whether exposure to noise has changed the population variance significantly (as an index of the extent to which there exists a real difference among individuals in susceptibility to traumatic noise deafness). Since the variances before and after have already been calculated for other purposes, it would be an enormous saving in time if the before and after variances could be used directly as an F ratio with any necessary allowance for correlation due to paired readings.

ANSWER: Since each ear has a measurement before and after exposure, the variances before and after may well be correlated. If so, the F test (described in this *Bulletin*, Vol. 1, page 70, query number 2) is not applicable because it rests on the assumption that the two estimates of variance are independent.

For variances based on correlated variates, the Pitman and Morgan method of testing the null hypothesis is described by Cochran in the same number of the *Bulletin* cited above, query number 3. Since this method requires calculation of r, you can save some labor by the following method of computation:

Before	After
x_1	x_1'
x_2	x_2'
.	.
.	.
.	.
x_k	x_k'
Sx	Sx'

Calculate the corrected sums of squares and products, Sx^2, $S(x')^2$, Sxx'. From these,

$$r=\frac{Sxx'}{\sqrt{Sx^2 \cdot S(x')^2}}$$

$$Sd^2=Sx^2+S(x')^2-2Sxx'$$

The correlation coefficient, r, is used in the Pitman and Morgan F test, while Sd^2 (the

corrected sum of squares of the differences, $x+x'$) enables you to complete the t test.

G. W. Snedecor

QUERY: In your answer to the first query of Vol. 1, page 70 of this *Bulletin*, you imply that the mean square group means should be divided by the mean square of individuals even though the quotient turns out to be less than 1.0. Have you decided not to use the rule, "Divide the larger mean square by the smaller"?

ANSWER: Two objectives must be distinguished. First, there is the reading of the table itself: since it contains no values less than 1.00, it can be entered only if F is greater than 1.00. This imposes the rule that goes with the table—divide the larger mean square by the smaller. There can be no deviation from this rule for using the table.

Second, there is the use to be made of the table, and this is varied. In answer to the query cited, the assumption was that the experimenter is interested in learning if his treatments serve to differentiate the population means. He is not concerned with a value of F less than that tabled for the probability which he has selected. Smaller values of F, including those less than 1.00, do not constitute evidence for the effectiveness of the treatments. For this experimenter, the rule for entering the table is always sufficient.

Another use for the table is the symmetrical or two-tailed test described in two answers contained in Vol. 1, pages 70-71, of the *Bulletin*. Again, the rule for entering the table is adequate, but the tabular probabilities must be doubled.

A third use of the table is the one that seems to be in querist's mind. He has mean squares for group means and for individuals, such as ordinarily arise in analysis of variance, but the former is the smaller. He evidently wishes to test the hypothesis that the mean square between groups is less than that within groups; that is, he wishes to make the asymmetrical test on the tail of the distribution with the small values of F. The nature of the distribution is such that this can be done easily: enter the table with the *reciprocal* of F (the larger variance divided by the smaller according to the rule), but *interchange the degrees of freedom*. For example, if there are three groups, each with nine individuals, and if F=0.10, enter the table with F=1/0.10 =10, n_1=24 and n_2=2, the 5% value being 19.45.

Several times I have observed significantly small values of F but have never found a reasonable interpretation. In every instance the small value seemed to be one of those unusual ones that occur occasionally in sampling from a homogeneous population. Have any readers of this column encountered a significantly small value of F with a realistic meaning?

G. W. Snedecor

QUERY: We have the problem of attempting to determine the contribution of the sire to rate of laying as measured by average clutch size. For example, 29 sires mated to 42 dams of one class with respect to clutch size gave 224 daughters. We have calculated the mean clutch size of the daughters of each sire. These means have been used to calculate a standard deviation that was weighted by using the frequency of daughters. From this a value of 0.4105 was obtained. The standard deviation of the population of 224 daughters was 1.0578. Using Pearl's formula for the correlation ratio (1940 edition, page 429) the result is 0.4105/1.0578=0.3881. Does this value measure the correlation between sires and daughters?

ANSWER: No. Although your data have a superficial resemblance to those considered by Pearl, they are fundamentally different. He was considering two independently measured variates with y-arrays in equally spaced x-intervals. You have only one variate, the average clutch size of daughters. The total sum of squares may be partitioned between sires and daughters as follows:

Source of variation	Degrees of freedom	Mean Square
Total	223	
Sires	28	S
Daughters of same sire	195	D

The total mean square is $(1.0578)^2$=1.1189

(assuming that you divided by degrees of freedom, 223), but I am unable to judge from your description what the mean square for sires is.

Assuming normal distribution in the sampled population, the completed analysis of variance leads to an estimate of *intraclass correlation*,

$$\frac{S-D}{S+(k_0-1)D},$$

in which k_0 is an average number of daughters per sire. If k represents the number of daughters for each sire, then

$$k_0=\frac{1}{28}\left(Sk-\frac{Sk^2}{Sk}\right)$$

Note: Subsequent correspondence elicited the information that S=1.348, D=1.092 and k_0=7.627. From these data the correlation 0.030 was calculated. The value is not significant, since F=1.348/1.092=1.23 whereas $F_{.05}$=1.53.

QUERY: We are planning an experiment with insecticides to be applied to sweet corn for the control of the European corn borer. The degree of infestation cannot be predicted with any degree of certainty. Tentative arrangements for the field experiments are as follows:

Three treatments (A, B and C) with check (X) are to be applied to 12 by 45 foot plots; that is, four rows each 45 feet long. The field plan is this:

A	X	B	X	C	X
X	B	X	A	X	C
C	X	B	X	A	X
X	A	X	C	X	B
B	X	A	X	C	X
X	C	X	B	X	A

The large number of check plots is necessary to compensate for drift of insecticide from one plot to another.

Samples are to be taken for the number of infested plants from 30 plants as near the center of each plot as can conveniently be done. The effect of the insecticide is to be evaluated on the basis of the number of borers, the yield and the quality of the corn.

All plots are to be planted to a variety of susceptible sweet corn, the planting to be as early as possible in order that a maximum infestation take place.

ANSWER: For the evaluation of yield your design is not very efficient. The desirable arrangement is one in which the plots of a replication lie as close together as feasible so as to avoid large soil differences among them, the replication (block) itself being nearly square. To avoid drifting of the insecticide and migration of the insects, greater separation is called for; hence, compromises have to be made.

Migration of the corn borer larvae is not a complication, I am told, but allowance must be made for drift. Perhaps it will be sufficient to spray only the two inner rows of your four-row plots, leaving the outer rows, as well as the ends of the inner, to absorb the drifting insecticide from adjacent plots. Absence of migration makes this plan available. Evaluations would be made on some 70 plants in a space about 6x35 feet in the central portion of the plot. This would leave at least 9 feet for the drift to subside. If that is not enough, another row could be put in between the plots, increasing the separation to 12 feet.

Such a design with four plots lying parallel would make the size of the replication 48x45 or 60x45, either of which would be satisfactory. There would be sufficient plants per plot to evaluate yield. If infestation by the borer is rather uniform, not all the plants would have to be dissected: the plots could be subsampled for damage determination. I am assuming that in sweet corn, yield and infestation can be measured on the same plants at harvest.

G. W. Snedecor

ABSTRACTS

(22)

STRANDSKOV, HERLUF H. and G. J. SIEMENS, (University of Chicago), **An Analysis of the Sex Ratios Among Single and Plural Births in the Total, the "White" and the "Colored" U. S. Populations. (To be published in the American Journal of Physical Anthropology, 3:–, 1946).**

Fisher's method of determining whether one variance is significantly greater than another is applied to certain U. S. Census data. The variances compared are the variances of percentages of male births over a 15 year period and the corresponding variances expected due to chance during the same period of time. For nearly all of the distributions which are examined the observed variance is found to be significantly greater than that expected due to chance.

The means of the percentages of males among single, twin, triplet and quadruplet births over a 15 year period are compared for significance of differences. It is found that the percentages of males decreases significantly with each increase in number of fetuses per pregnancy or as the mammalogist would say with each increase in size of litter.

Racial differences in the percentage of males among the different types of births are tested and found generally to be significantly different.

(23)*

WADLEY. F. M. (USDA Bureau of Entomology nd Plant Quarantine), **Incomplete Block Design Adapted to Paired Tests of Mosquito Repellents.**

The application of such a plan is described. There can be only two arms ("plots") in the "block", which is a subject on a given date. Analysis followed Yates' standard method. The scheme worked well and gave a worthwhile gain in efficiency, compared to the alternate plan of using a standard in each pair. It will probably be of value in such situations.

(24)

ANDERSON, R. L. (Institute of Statistics). **The Analysis of Orthogonal Square Lattice Experiments with *d* duplications of the Basic Design.**

A generalized method of analyzing any square lattice experiment with k^2 treatments put in blocks of k each is presented, provided the r replications in the basic designs are orthogonal. The basic design is duplicated d times. For a 6x6 triple lattice using 9 complete replications, k=6, r=3, and d=3.

The methods of analysis follow those developed by Yates and Cochran, utilizing any added information contained in the inter-block variance. Yates has denoted the weighting factor for making block adjustments to be

$$k\mu=\frac{w-w'}{(r-1)\,w+w'}$$

where $\frac{1}{w}=\sigma_e^2$, the true intra-block error, and

$\frac{1}{w'}=\sigma_e^2+k\sigma_b^2$, the true inter-block error.

It is shown that the best estimates of w and w' are (respectively)

$$\frac{1}{E_e} \text{ and } \frac{rd-1}{rdE_b-E_e}$$

where E_e is the computed intra-block variance and E_b the computed inter-block variance. E_b is found by pooling components (a) and (b) in the analysis of variance.

The average variance of the difference between two adjusted treatment means is

$$\frac{2}{wdr}\left[1+\frac{rk}{k+1}\,\mu\right]$$

(25)

MUHRER, M. E. and A. G. HOGAN. (University of Missouri). **Effect of Goitrogenic Drugs on Fattening Swine. Proc. Soc. Exp. Biol. Med. 60:211-212. 1945.**

Studies were made of the effect of thiouracil and thiourea upon the economy of gain and morphology (body measurements) in swine. The thiouracil-fed animals made greater gains than either the thiourea-fed animals or the animals which received the basal ration alone. In the analysis of variance in gains by the method of Snedecor the value of F (ratio of the standard deviation squared between and within treatments) was 23.6 and greatly exceeds the 1% point. Along with other body measurements the increase in height and depth were studied in relation to the increase in weight. From the analysis of variance it was

*Biometrics Bulletin, Vol. 2, No. 2, page 30. April 1946.

found that the ratio of the weight increase to height was more significant than either weight or height increase. However, the ratio of weight increase to depth increase was not significant. The thiouracil kept the animals significantly shorter than the control animals but affected the depth increase only through affecting the size of the animal.

NEWS AND NOTES

U. S. NAIR, Head of the Department of Statistics, Travancore University, Trivandrum, South India, writes, "You will be interested in knowing that a Statistical Laboratory has been created in our University where post graduate tuition in statistics is given; also, research fellows work in the Laboratory. The post graduate course consists of two years during which time intensive training in mathematical analysis, theoretical and practical statistics will be given. The subjects in applied statistics include design of experiments, factor analysis, statistical physics, biometry and certain aspects of mathematical economics." He urges us to send publications which will be helpful to him, "living as we do in one of the remotest corners of the world." . . . Word has been received from V. G. PANSE, Institute of Plant Industry, Indore, Central India. He has published an account of sample surveys carried out for estimating the yield of commercial crops of cotton . . . D. D. KOSAMBI, Professor of Mathematics, is with the Tata Institute of fundamental research, Bombay 26, India . . . The Proceedings of the 33rd Indian Science Congress, Bangalore 1946 shows a very active Section on Statistics. K. B. MADHAVA is president of the Section. He is a University Professor of Mysore who is now on foreign service with the government of India, Transport Department, New Delhi. The title of his Presidential Address was "Statistics gets firmly woven into our fabric of thinking"—a most interesting plea for statistics. He writes, "through sheer power of logic, the statistical method has secured a place for itself in all fields of thought." There were 47 papers listed in the report of the Abstracts of papers discussed at the Congress. The papers were grouped under theoretical, agricultural, economic, vital and general statistics. P. K. BOSE and P. C. MAHALANOBIS, Presidency College, Calcutta, and A. R. SEN, Lucknow were the speakers in the agricultural group. In the vital statistics section, C. CHANDRA SEKAR, Calcutta, reported on "Reproductive wastage and infant mortality as obtained from the records of some maternity and Child Welfare centers in Calcutta." Another paper was by K. K. MATHEN and R. B. LAL, Calcutta, dealing with "Studies in the health problems of a rural community in western Bengal. Part I. Population problems" . . . The Deputy Director of Agriculture, (Crop Research) Bombay Province is V. M. CHAVAN, College of Agriculture, Poona . . . P. V. SUKHATME is Statistical Adviser, Imperial Council of Agricultural Research, New Delhi . . . R. J. KALAMKAR is an Officer on Special Duty, Office of the Director of Agriculture, Central Provinces, Nagpur . . . R. C. BOSE is Head of the Postgraduate Department of Statistics, Calcutta University, Calcutta, India . . . B. M. PUGH, Professor of Agronomy at Allahabad Agricultural Institute and Editor of the Allahabad Farmer intends to leave India in July to visit the United States. He will be visiting the agricultural colleges of this country . . . JOSEPH CARMIN sends an announcement regarding the Independent Biological Laboratories, Kefar-Malal, Ramatayim, Palestine. "This announcement is intended to let us know that we went safely through the war, that we are continuing our work, and that we are trying now to get in touch anew with institutions and scientists abroad." They were forced to leave their precious domicile at Tel-Aviv and move in great haste to the country. Buildings have been erected now to house their library, collections, apparatus and 12 research workers. "Research was going on uninterruptedly all the time of the war on utilization of plant and animal material for the manufacture of

different commodities and a special consulting bureau was established in this line. Work was continued on previous lines in different problems of ecology, genetics and physiology of plants and animals, entomology, phytopathology and marine biology" . . . Other biological research workers that we have heard from recently are P. S. OSTERGAARD who is with the Agricultural Research Laboratory, Copenhagen, Denmark; HALVDAN ASTRAND, chief statistician of the Swedish Sugar Company, Arlov; S. BERGE, professor of animal breeding and genetics at the Agricultural College of Norway, Aas; A. SCIUCHETTI also in the field of animal breeding and genetics at the Agriculture School, Plantahol, Landquart (Switzerland); and S. H. JAYEWICKREME, Division of Medical Entomology, Colombo, Ceylon . . . ALVIN KEZER, chief agronomist at Colorado A. & M. College lets us know about the Diamond Anniversary of his school . . . K. M. AUTREY is now Associate Professor of Dairy Husbandry at The Pennsylvania State College . . . Lt. Colonel JAMES H. BYWATERS returned to the U. S. Regional Laboratory at East Lansing, Michigan, early in May. While on terminal leave, he and Mrs. Bywaters visited J. HOLMES MARTIN, Head of Department of Poultry Husbandry, Purdue University. Think they also visited Iowa State College . . . WALTER C. JACOB, recently Lt. Commander with the Bureau of Ships Statistics Section, has just resumed his former connections with Cornell University and is located at the Long Island Vegetable Research Farm near Riverhead, Long Island . . . CHARLES M. MOTTLEY has "reverted to my old field of work in fishery biology." He is Chief of the Section of Eastern Agricultural Investigations in the Interior Department . . . RUTH R. PUFFER, director of Statistical Service of the Tennessee Department of Public Health, has been granted a leave of absence to serve as a visiting professor in the School of Public Health of the University of Chile for the term, June-August, 1946. In addition to her duties in connection with the University, she will be a consultant on statistical and tuberculosis work for Chile. She has gone to Chile under the sponsorship of the Rockefeller Foundation, which has been instrumental in the establishment of this School of Public Health . . . FRANK A. WECK was a Captain with the Office of the Surgeon General. He is with the Acturial Division, Metropolitan Life Insurance Company, New York . . . CARL F. KOSSACK who was in the Department of Mathematics at the University of Oregon has recently been appointed Mathematician with the Joint Army-Navy Air Intelligence Division.

Officers of the American Statistical Association: President, Isador Lubin; Directors, Chester I. Bliss, E. Grosvenor Plowman, Walter A. Shewhart, Samuel A. Stouffer, Willard L. Thorp, Helen M. Walker; Vice-Presidents, F. L. Carmichael, S. S. Wilks, Dorothy Swaine Thomas; Secretary-Treasurer, Lester S. Kellogg.

Officers of the Biometrics Section: Chairman, D. B. DeLury; Secretary, H. W. Norton; Section Committee members; E. J. deBeer, A. E. Brandt, J. W. Fertig, J. G. Osborne, J. W. Tukey.

Editorial Committee for the Biometrics Bulletin: Chairman, Gertrude Cox; members, R. L. Anderson, C. I. Bliss, W. G. Cochran, Churchill Eisenhart, H. W. Norton, G. W. Snedecor, C. P. Winsor.

Material for the BULLETIN should be addressed to the Chairman of the Editorial Committee, Institute of Statistics, North Carolina State College, Raleigh, N. C., material for Queries should go to "Queries," Statistical Laboratory, Iowa State College, Ames, Iowa, or to any member of the committee.